Assignment 3

3.1 Def^n: a is a subgradient of f at x if

$$f(y) \geq f(x) + a^T(y-x) \quad \forall y$$

Group 3: Good!

* Similarly, since the dual function g is concave, a is a supergradient of g at x if

$$g(y) \leq g(x) + a^T(y-x) \quad \forall y.$$

Group 3: Might be good with some elaboration, such as by mentioning that f is convex and that g is the dual function corresponding to f.

$$\text{minimize} \quad f(w)$$
$$\text{s.t} \quad Aw = b$$

Then; $g(\lambda) = \inf_w L(w,\lambda) = f(w) + \lambda^T(Aw-b)$

Then, for $\lambda_0 \in \mathbb{R}^n$,

$$g(\lambda_0) = \inf_w f(w) + \lambda_0^T(Aw-b)$$

Let $w_{\lambda_0} = \underset{w}{\operatorname{argmin}}\ f(w) + \lambda_0^T(Aw-b)$

$$\Rightarrow g(\lambda_0) = f(w_{\lambda_0}) + \lambda_0^T(Aw_{\lambda_0}-b) \quad \text{—— (1)}$$

Similarly, ~~let~~ for any $\lambda \in \mathbb{R}^n$ let

$$w_\lambda = \underset{w}{\operatorname{argmin}}\ f(w) + \lambda^T(Aw-b)$$

$$\Rightarrow g(\lambda) = f(w_\lambda) + \lambda^T(Aw_\lambda - b) \quad \text{—— (2)}$$

Note that

$$g(\lambda) = \inf_{w} f(w) + \lambda^T(Aw - b)$$

$$\leq f(w) + \lambda^T(Aw-b) \quad \text{for all } w \quad (3)$$

$\therefore$ (3) is true for $w = w_{\lambda_0}$

$\Longrightarrow$

$$g(\lambda) \leq f(w_{\lambda_0}) + \lambda^T(Aw_{\lambda_0} - b) \quad \text{—— } (4)$$

Then, substituting $f(w_{\lambda_0})$ using (1) in (4)

gives

$$g(\lambda) \leq g(\lambda_0) - \lambda_0^T(Aw_{\lambda_0} - b) + \lambda^T(Aw_{\lambda_0} - b)$$

$$= g(\lambda_0) + (\lambda - \lambda_0)^T(Aw_{\lambda_0} - b)$$

$$= g(\lambda_0) + (Aw_{\lambda_0} - b)^T(\lambda - \lambda_0)$$

$\Longrightarrow (Aw_{\lambda_0} - b)$ is a supergradient of g at λ_0

$\therefore (Aw - b) \in \partial g(\lambda)$

$\uparrow$ set of supergradients of g at λ

Group 3: Nice work!

3.2 ~~Since f is smooth and strongly convex, g is also smooth and strongly convex~~

Group 3: Good that you mentioned f is L-smooth and mu-strongly convex! Better too detailed than unclear :)

Since f is L-smooth and μ-strongly convex, g is also smooth and strongly concave with constants $L_g = \frac{\lambda_{max}(AA^T)}{\mu}$ and

$\mu_g = \frac{\lambda^+_{min}(AA^T)}{L}$, respectively, where

λ_{max} is the maximum eigenvalue of AA^T and λ^+_{min} is the minimum nonzero eigenvalue of AA^T

(Above is a well-known theorem)

* Let $l(\lambda) = -g(\lambda)$ (the negative dual fu^n)

Then $\nabla l(\lambda) = -\nabla g(\lambda)$

Since l is L_g smooth, we have

$$l(x_2) \le l(x_1) + \nabla l(x_1)^T (x_2 - x_1) + \frac{L_g}{2} \|x_2 - x_1\|^2 \quad (1)$$

Let $x_2 = \lambda^{k+1}$ & $x_1 = \lambda^k$. Then

$$x_2 - x_1 = \lambda^{k+1} - \lambda^k = -\alpha_k \nabla l(\lambda^k),$$

$$(\because \lambda^{k+1} = \lambda^k - \alpha_k \nabla l(\lambda^k))$$

$\therefore$ (1) $\Longrightarrow$

$$l(\lambda^{k+1}) \le l(\lambda^k) + \nabla l(\lambda^k)^T (-\alpha_k \nabla l(\lambda^k)) + \frac{L_g}{2} \alpha_k^2 \|\nabla l(\lambda^k)\|^2$$

$$\Longrightarrow l(\lambda^{k+1}) \le l(\lambda^k) - \alpha_k \| l(\lambda^k) \|^2 + \frac{L_g}{2} \alpha_k^2 \| \nabla l(\lambda^k) \|^2$$

$$\le l(\lambda^k) - \alpha_k \| l(\lambda^k) \|^2 + \frac{L_g}{2} \alpha_k \times \frac{1}{L_g} \| \nabla l(\lambda^k) \|^2$$

for all $\alpha_k \le \frac{1}{L_g}$

$$= l(\lambda^k) - \frac{\alpha_k}{2} \| \nabla l(\lambda^k) \|^2 \quad \text{—— (2)}$$

Since $l(\lambda)$ is strongly convex with constant m_g, we have that

$$\| \nabla h(\lambda^k) \|^2 \ge 2 m_g (l(\lambda^k) - l(\lambda^*))$$

Then (2) $\Longrightarrow$

$$l(\lambda^{k+1}) \le l(\lambda^k) - \frac{\alpha_k}{2} \times 2 m_g (l(\lambda^k) - l(\lambda^*))$$

$$= (1 - \alpha_k m_g) l(\lambda^k) + \alpha_k m_g l(\lambda^*)$$

$$\Longrightarrow l(\lambda^{k+1}) - l(\lambda^*) \le (1 - \alpha_k m_g)(l(\lambda^k) - l(\lambda^*)) \quad (3)$$

Let $\alpha_k = \alpha$ (constant stem size).

Then by using recursive application of (3) we get

$$l(\lambda^k) - l(\lambda^*) \le (1 - \alpha m_g)^k [l(\lambda^0) - l(\lambda^*)] \quad (4)$$

Group 3: Good use of equation references!

~~Here $1-\alpha r_g$~~

Here $0 < 1 - \alpha r_g < 1$, since $0 < \alpha < \frac{1}{L_g}$ and $r_g \leq L_g$.

$\therefore$ (4) $\Longrightarrow$ $h(\lambda^k)$ converses to the optimal value $h(\lambda^*)$ with a linear rate.

Group 3: Nice!

* The solution $w_{k+1} = \arg\min_{w} (L w, \lambda_k)$ is primal freasible only when the algorithm converses to it's optimal value. In the intermediate itterations, the solution w_{k+1} is not ~~prima~~ feasible.

Group 3: Would be great if you can elaborate this a little bit.

3.3 (P_2): minimize $\frac{1}{N} \sum f_i(w_i)$

s.t $w_i = w_j \quad \forall j \in N_i \;, i \in [N]$

~~Write P_2 equivalently as~~

~~minimize $\frac{1}{N} \sum_{i \in [N]} f_i(w_i)$~~

~~s.t $a_{ij}(w_i - w_j) = 0 \quad \forall j \in N_i, \; i \in [N]$,~~

~~Where $A = [a_{ij}]$ is a doubly stochastic matrix.~~

Group 3: I suggest using the mobile app Scannable for better images.

Write (P_2) equivalently as

$$\text{minimize} \quad \frac{1}{N}\sum_{i\in[N]} f_i(w_i)$$

$$\text{s.t} \quad a_{ij}(w_i - w_j) = 0 \quad \forall\, j \in N_i \quad \text{, where}$$

$A = [a_{ij}]$ is a doubly stochastic matrix compatible with an arbitrary undirected and connected graph.

Then adding all the constraints associated with each $j \in N_i$ gives us

$$\sum_{j\in N_i} a_{ij}(w_i - w_j) = 0$$

$$\Longrightarrow \sum_{j\in N_i} a_{ij} w_i = \sum_{j\in N_i} a_{ij} w_j$$

$$\Longrightarrow \underbrace{\left(\sum_{j\in N_i} a_{ij}\right)}_{=1} w_i = \sum_{j\in N_i} a_{ij} w_j$$

$$\Longrightarrow w_i = \sum_{j\in N_i} a_{ij} w_j \quad \forall\, i \in [N]$$

Then (P_2) can reformulate as

$$\underset{w}{\text{minimize}} \quad \frac{1}{N}\sum_{i\in N} f_i(w_i)$$

$$\text{s.t} \quad w_i - \sum_{j\in N_i} a_{ij} w_j = 0 \quad ; \; i \in [N]$$

Then ~~$g(\lambda)$~~ the dual fn g of $(P2)$ is

$$g(\lambda) = \inf_{W} \left[\sum_{i \in [N]} \frac{f_i(w_i)}{N} + \sum_{i \in [N]} \lambda_i^T \Big(w_i - \sum_{j \in N_i} a_{ij} w_j \Big) \right]$$

$$= \inf_{W} \left[\sum_{i \in [N]} \left\{ \frac{f_i(w_i)}{N} + \lambda_i^T \Big(w_i - \sum_{j \in N_i} a_{ij} w_j \Big) \right\} \right]$$

$$= \sum_{i \in [N]} \underbrace{\inf_{w_i} \left(\frac{f_i(w_i)}{N} + \lambda_i^T \Big(w_i - \sum_{j \in N_i} a_{ij} w_j \Big) \right)}_{\text{subproblems.}}$$

$$\Big(w_i - a_{ii} w_i - \sum_{j \in N_i \setminus \{i\}} a_{ij} w_j \Big)$$

Then the corresponding ~~dual~~ distributed dual ascent algorithm is

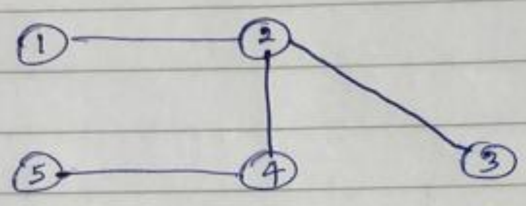
$$\cancel{w_i^{(k+1)} = \operatorname*{argmin}_{w_i} \frac{f_i(w_i)}{N} + \left(\lambda_i^T\right)^{(k)} w_i - \sum a_{ij} w}$$

primal variable update $\Rightarrow$ $$w_i^{(k+1)} = \operatorname*{argmin}_{w_i} \frac{f_i(w_i)}{N} + \left(\lambda_i^T\right)^{(k)} \Big[(1 - a_{ii}) w_i - \sum_{j \in N_i - \{i\}} a_{ij} w_j^{(k)} \Big]$$

dual variable component update $\Rightarrow$ $$\lambda_i^{(k+1)} = \lambda_i^{(k)} + \alpha_k \Big[(1 - a_{ii}) w_i^{(k+1)} - \sum_{j \in N_i - \{i\}} a_{ij} w_j \Big]$$

Comparison between the dual method and the primal method (numerically) ~~for a~~ using a particular connected and undirected graph.

We consider the following communication graph with 5 users.



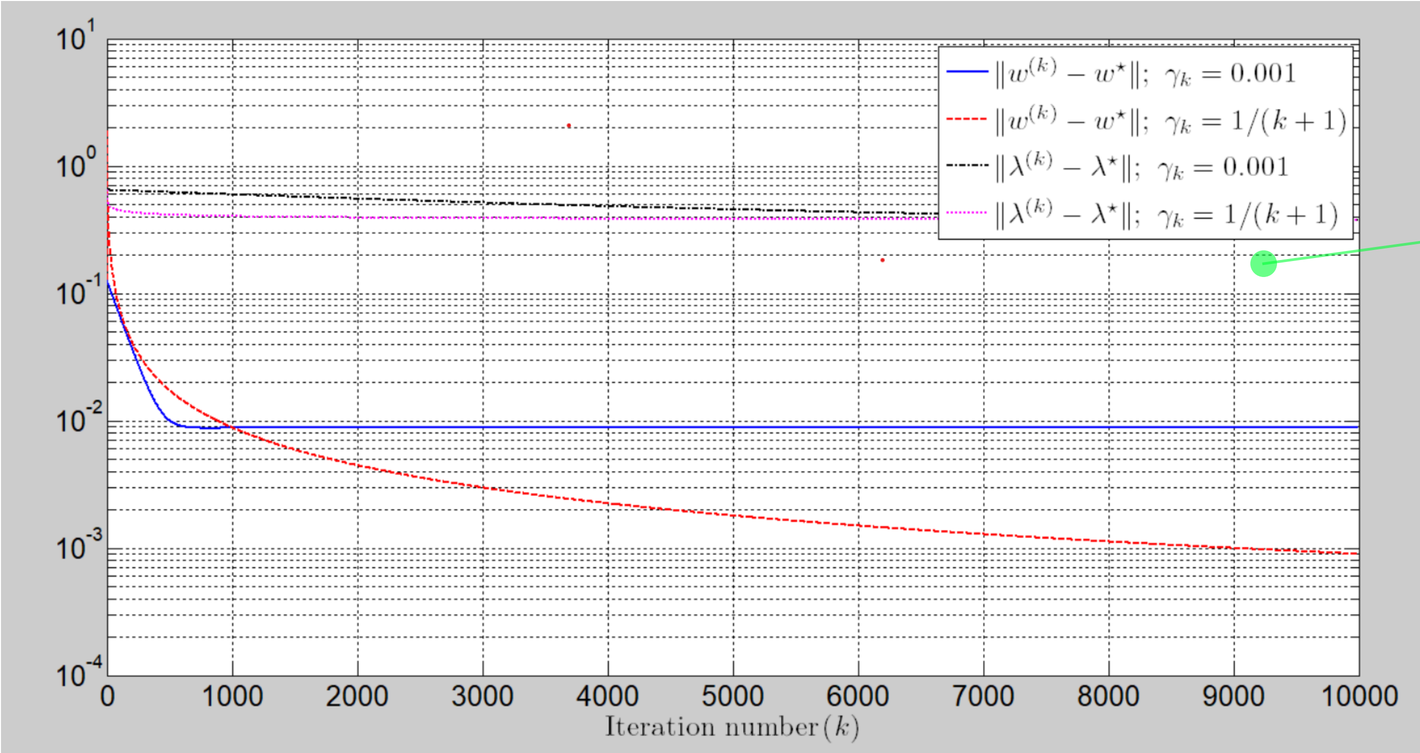
We used:

* $f_i(w_i) = w_i^T B_i w_i + q_i^T w_i$, ~~where~~ $i = 1, 2, 3, 4, 5$,
  where B_i s are positive definite matrices.
* $w_i \in \mathbb{R}$
* Doubly stochastic matrix A is taken as

$$A = \begin{bmatrix} 0.75 & 0.25 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 & 0 \\ 0 & 0.25 & 0.75 & 0 & 0 \\ 0 & 0.25 & 0 & 5/12 & 1/3 \\ 0 & 0 & 0 & 1/3 & 2/3 \end{bmatrix}$$

Following figure shows the convergences of $\|w^{(k)} - w^*\|$ and $\|\lambda^{(k)} - \lambda^*\|$ using the primal method and the dual method, respectively.

Group 3: Would be great if you could add the code here.



Group 3: Would be great if you can add the name of each method in the legend!

According to the graph, for this particular case, the primal method shows better convergences than the dual method. For constant step size, the primal method ~~shows~~ clearly shows a linear rate of convergence. However, with constant stepsize, $w^{(k)}$ seems to converge into a neighbourhood of w^*, while with nonsummable and square summable stepsize ~~(1/k+1)~~ $(\alpha_k = 1/k+1)$ $w^{(k)}$ converges more towards the optimal solution w^*.

In both methods, at every iteration, each ~~the~~ users ~~communication with~~ communicate only with it's neighbours. $\therefore$ the communication cost in both methods ~~per iteration~~ are same, per iteration.

Group 3: Good work in overall! :)